Predictive Model of CO_2 Based on World Development Indicators

Objectives

Carbon dioxide (CO_2) is a greenhouse gas and major contributor to climate change. CO_2 emissions levels differ widely between countries (Figures 1, 2). As human activities are a major source of CO_2 emissions, this project aimed to understand the economic, environmental, and technological factors contributing to CO_2 emissions. Specifically, using World Bank World Development Indicators data, we built a predictive model of CO_2 emissions per capita based on countries' gross domestic product (GDP), economic activities contributing to GDP (e.g. Agriculture, Industry, Services, Manufacturing), environmental factors (e.g. forest area), and technologies (e.g. access to clean cooking fuels, access to electricity).

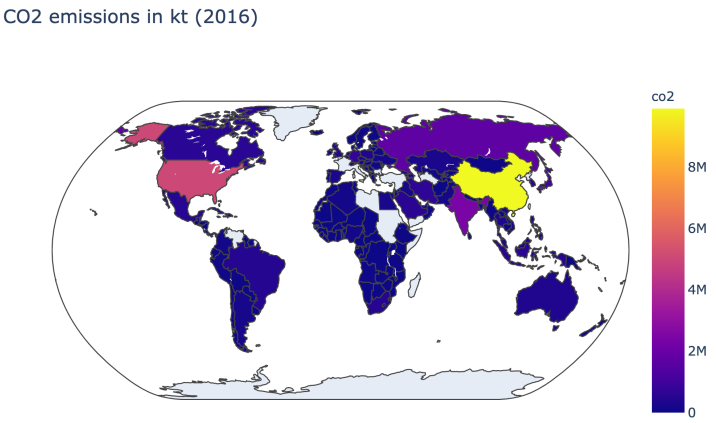


Figure 1. CO_2 emissions levels (in kt) by country, from the World Bank "World Development Indicators" dataset. See *Data Preparation* below for dataset description.

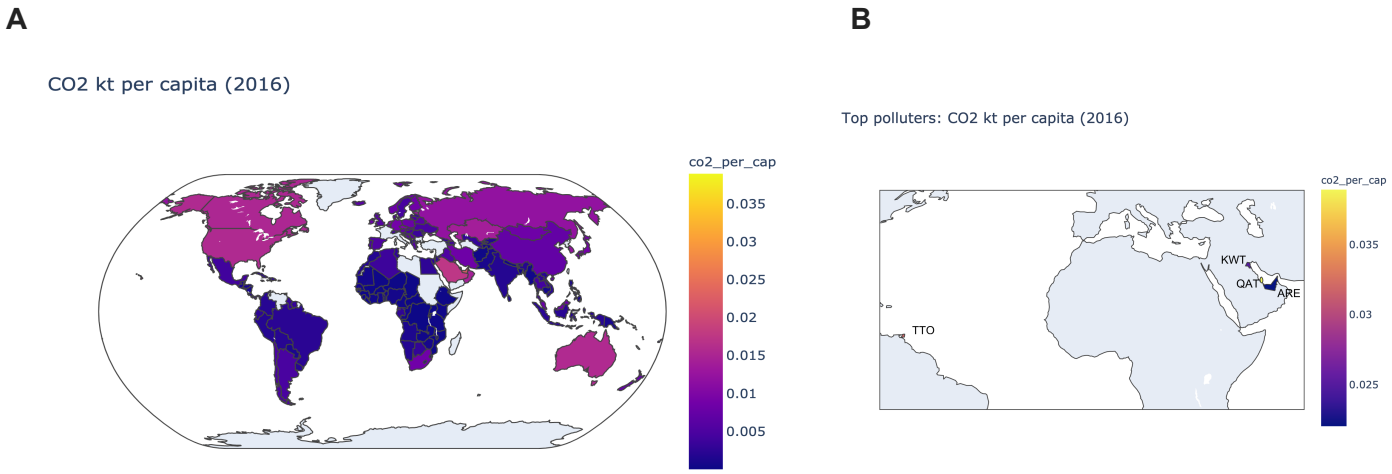


Figure 2. (A) CO_2 emissions per capita (in kt per capita) by country, from the World Bank "World Development Indicators" dataset. **(B)** Zoomed in view of top polluters per capita: ARE (United Arab Emirates), KWT (Kuwait), QAT (Qatar), TTO (Trinidad and Tobago). See *Data Preparation* below for dataset description.

Data Preparation

Data source: The World Bank DataBank

The World Bank is an organization dedicated to reducing poverty, increasing shared prosperity, and promoting sustainable development. They provide transparent, accurate and high quality data that is open source. The World Bank Development Indicator data was chosen from DataBank, the world bank database which contains millions of records with a large amount of variables, countries, and years to choose from. The flexibility of the data allowed a more diverse and precise analysis for the project. The variables chosen for this project were the human economic, environmental, and technological activities believed to have the biggest effect on CO_2 emissions per capita. The database contained data from 40 years but 2016 was chosen because it was one of the most recent years that had the least amount of null values in the dataset. Choosing one year with strong data will allow us to find a strong correlation between the different predictor variables and CO_2 emissions per capita.

Data cleaning and feature engineering

An analysis was done to determine the best method to handle the null values within the predictors.

The methods explored were either replacing the null values with the median or dropping the null values. It was determined that dropping the null values gave more accurate results when looking at the OLS regression.

Initially, the y variable was going to be CO_2 emissions but further analysis showed that the CO_2 emissions (in kt) per capita (CO_2/pop) showed stronger relationships with the X variables (Figure 3). As all of the X values relate closely to GDP, it makes the population of the country important because there is a strong relationship between the number of people from a country that contribute to the GDP and CO_2 emissions. Therefore, using CO_2 emissions per capita as the y variable provides more accurate results during the analysis.

Variable descriptions and transformations

To understand the effects of economic, environmental, and technological factors on CO_2 emissions per capita, we selected the predictors from the World Bank Development Indicator database related to GDP, economic activities and industries contributing to GDP, forest area, electricity access, and clean cooking technology (Table 1). We believe these variables contribute and cover many aspects of CO_2 emissions within different countries. We performed variable transformations (Table 1) to satisfy the assumption of linearity between X and y (Figure 3).

Table 1. Variables included in the model and transformations applied prior to modeling.

Feature	Variable code	X or y variable	Transformation
CO_2 emissions (in kt) per capita	co2_per_cap (= co2 / pop)	y	log10
GDP per capita (current US$)	gdp_per_cap	X	log10
GDP growth (annual %)	pgdp_growth	X	Yeo-Johnson
Agriculture, forestry, and fishing, value added (% of GDP)	pgdp_nat	X	Square Root
Industry (including construction), value added (% of GDP)	pgdp_ind	X	Square Root

Manufacturing, value added (% of GDP)	pgdp_man	X	Square Root
Services, value added (% of GDP)	pgdp_ser	X	None
Forest area (% of land area)	pland_forest	X	None
Population, total	pop	X	Box-Cox
Access to clean fuels and technologies for cooking (% of population)	ppop_cfuel	X	Square Root
Access to electricity (% of population)	ppop_electric	X	Square Root

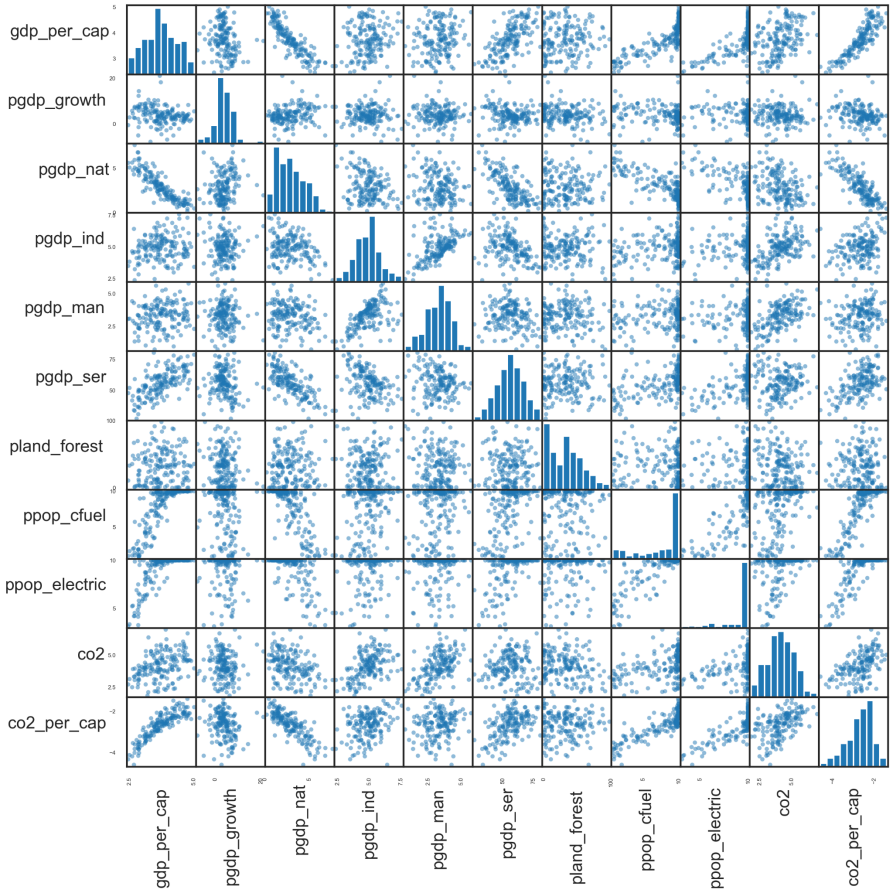


Figure 3. Scatter plot matrix of transformed variables (see Table 1 for transformations performed). The y variable for the remainder of this analysis was CO_2 in kt per capita (co2_per_cap).

Back-transformation methods

We used the following back-transformation procedures to interpret the model slope and intercept in terms of untransformed CO_2 emissions in kt per capita and untransformed X variables, where **intercept$_{tr}$, slope$_{tr}$**, and **X$_{tr}$** represent a **transformed intercept, slope,** or **X variable**, respectively:

- *log10:* 10**intercept$_{tr}$, 100∗(10**(slope$_{tr}$)−1) %
- *Square root:* X$_{tr}$**2
- *Box-Cox:* scipy.special.inv_boxcox(X$_{tr}$, lmbda)

Correlation analysis

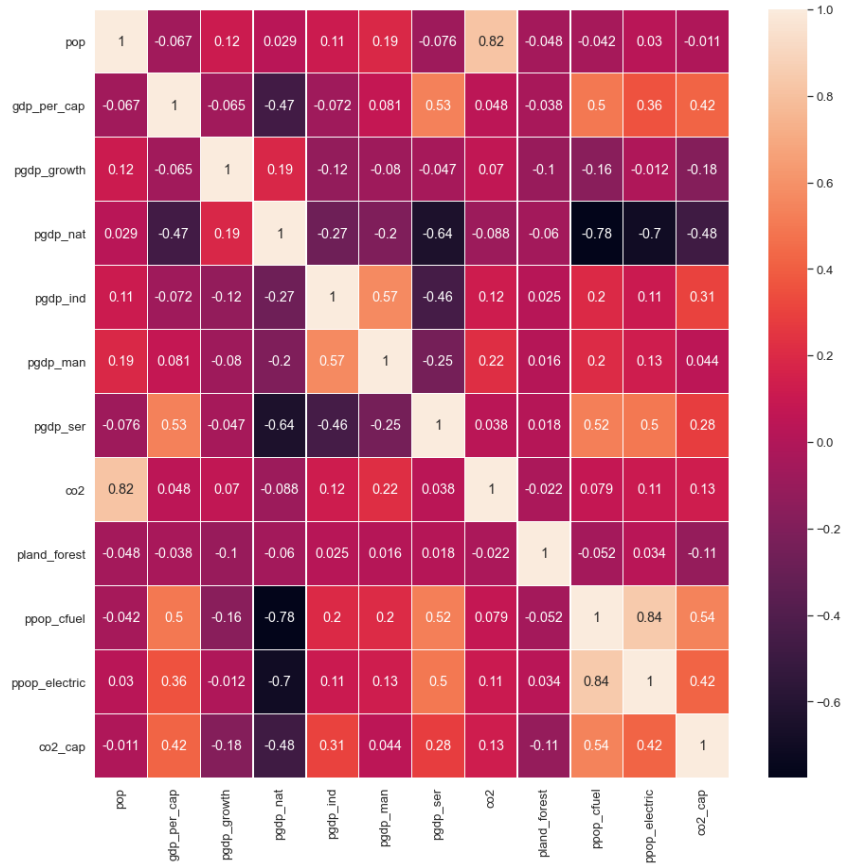


Figure 4. Heatmap diagram of correlation between raw data (Pearson correlation coefficients).

As expected from common social studies, population and the amount of CO_2 indicated strong positive correlation. Due to the similarity in nature, ppop_electric and ppop_cfuel also indicated strong positive correlation. Strong negative correlations could be found for the following pairs: ppop_cfuel *vs.* pgdp_nat, ppop_electric *vs.* pgdp_nat, and pgdp_ser *vs.* pgdp_nat. The coefficients of these pairs indicated a possible problem of multicollinearity. However, by setting Variation Inflation Factor (VIF) < 10 as our threshold, we could assume that the correlation coefficients that are below 0.95 are acceptable in our analysis. Thus, we didn't exclude any predictors for our model.

Model

Rationale for model selection

Ordinary least squares (OLS) linear regression was well suited to our analysis for the following reasons:

- Our objective was to estimate the value of a random variable, CO_2 emissions per capita.
- All variables were continuous numeric variables, and the relationship between the transformed y and X variables was linear (Figure 3).

Simple linear regression models for each X variable

To better understand the relationship of each X variable (Table 1) with CO_2 emissions per capita, we first performed simple OLS linear regressions for each X variable. The intercept, slope, and

R-squared (fit) of each simple OLS model is summarized in Table 2. A brief summary of each model follows.

Table 2. Summary of key parameters from simple OLS linear regression models of log10 CO_2 emissions in kt per capita by each transformed X variable.

X variable in simple OLS (transformed as described in Table 1)	Intercept	Slope	R-squared
GDP per capita (current US$)	-11.66	3.97 (X) -0.41 (X^2)	0.813
GDP growth (annual %)	-2.49	-0.038	0.063
Agriculture, forestry, and fishing, value added (% of GDP)	-1.76	-0.36	0.633
Industry (including construction), value added (% of GDP)	-3.33	0.14	0.053
Manufacturing, value added (% of GDP)	-7.11	0.28	0.035
Services, value added (% of GDP)	-6.00	-0.0033	0.017
Forest area (% of land area)	-2.66	-0.0001	0.000
Access to clean fuels and technologies for cooking (% of population)	-4.02	0.18	0.742
Access to electricity (% of population)	-5.27	0.29	0.645
Population, total (in a model of CO_2 emissions in kt)	0.41	0.12	0.680

Population, total

Since CO_2 per capita is being used as a dependent variable, the relationship between the $\log_{10}$ transformed CO_2 and **Population (pop)** was analyzed. To perform simple OLS analysis, Box-Cox Transformation was performed on population. The OLS model explained 68.0% of the variance (R-squared = 0.680), indicating relatively good fit.

GDP per capita (current US$)

A log base 10 transformation was used on **GDP per capita (gdp_per_cap)**. However, the OLS model of log10_co2_per_cap by log10_gdp_per_cap showed a parabola shape in the residuals *vs.* predicted values and residuals *vs.* X plots. It indicated a lack of fit in the model and it suggested the need for a quadratic term in the model. To fix this, a quadratic term was added and the OLS model was re-fitted. As a result, the residuals vs. predicted values and residuals vs. X plots were fixed. This quadratic regression showed that 81.3% of variation was explained by only the independent variables that actually affect the dependent variable (i.e. adjusted R-squared = 0.813).

GDP growth (annual %)

Given there is negative value in the GDP growth variable, a Yeo-Johnson transformation was used on **GDP growth (pgdp_growth)**, instead of the Box-Cox transformation. An OLS model of log10_co2_per_cap by yj_pgdp_growth explained 6.3% of the variance in the target variable (R-squared = 0.063), indicating a poor fit.

Agriculture, forestry, and fishing, value added (% of GDP)

A square root transformation was used on **Agriculture, forestry, and fishing, value added (% of GDP) (pgdp_nat_sqrt)**. An OLS model of log10_co2_per_cap by pgdp_nat_sqrt explained 63.3% of the variance in the target variable (R-squared = 0.633), indicating relatively good fit. The slope of the model was negative, indicating that as pgdp_nat increases, co2_per_cap decreases.

Industry (including construction), value added (% of GDP)

A square root transformation was used on **Industry (including construction) value added (% of GDP) (pgdp_ind_sqrt)**. An OLS model of log10_co2_per_cap by pgdp_ind_sqrt explained 5.3% of the variance in the target variable (R-squared = 0.053), indicating poor fit. The slope of the model was positive, indicating that as pgdp_nat increases, co2_per_cap increases.

Manufacturing, value added (% of GDP)

A square root transformation was used on the **Manufacturing value added (% of GDP as sqrt_pgdp_man)** dataset as that gave a normal distribution. An OLS model of log10_co2_per_cap by sqrt_pgdp_man was created to find the level of variance which was at 3.5% (R-squared = 0.035) indicating that this is a poor fit

Services, value added (% of GDP)

No transformations were performed on **Services value added (% of GDP as pgdp_ser)** because the dataset is already normally distributed. An OLS model of log10_co2_per_cap by pgdp_ser was created to find the level of variance which was at 1.7% (R-squared = 0.017). A low R-squared indicates that this is a poor fit.

Forest area (% of land area)

Various transformation methods on the **Forest Area (pland_forest)** didn't essentially change the nature of data at all; thus, no transformation was performed for simple analytic purposes. The OLS model of $\log_{10}$ transformed CO_2 per capita and Forest Area indicated very poor fit (R-squared = 0.000) with close to zero slope (slope = -0.0001). Therefore, it is safe to say that the relationship between the Forest Area and CO_2 per capita cannot be approached via linear models for the given dataset.

Access to clean fuels and technologies for cooking (% of population)

A square root transformation was used on **Access to clean fuels and technologies for cooking (% of population) (ppop_cfuel_sqrt)**. An OLS model of log10_co2_per_cap by ppop_cfuel_sqrt explained 74.2% of the variance in the target variable (R-squared = 0.742), indicating a relatively good fit. The slope of the model was positive, indicating that as ppop_cfuel increases, co2_per_cap increases.

Access to electricity (% of population)

A square root transformation was used on **Access to electricity (% of population) (ppop_electric_sqrt)**. An OLS model of log10_co2_per_cap by ppop_electric_sqrt explained 64.5% of the variance in the target variable (R-squared = 0.645), indicating relatively good fit. The slope of the model was positive, indicating that as ppop_electric increases, co2_per_cap increases.

Multiple linear regression model

Dropped Variables (Part 1)

CO_2 emissions per capita is used as the dependent variable instead of CO_2 emissions by country (see *Data Preparation*). The predictor variable "population" is, therefore, dropped after used to calculate CO_2 emissions per capita.

Initial Regression Model

Linearity was found between the dependent variable and predictors through simple OLS linear regression examinations (Table 2, Figure 3). The same transformations were applied to each predictors in the initial multi linear regression formula (ref. Table 1).

After fitting the model, an outlier, idx.44 - Congo (Fig.4) , was identified. Its associated residual is a clear outlier in all Residuals vs Predictor charts, and its related studentized residual reaches -4. These indicate that Cogno is an influential outlier. A deeper investigation suggests that Congo's CO_2 emission per capita is unusually low compared to the importance of the industry and manufacturing sector in its GDP (41.2% and 18.3% respectively). This is a unique case, likely caused by reasons not included within our data, *e.g.* geographic condition. We decided to exclude this outlier in our data set.

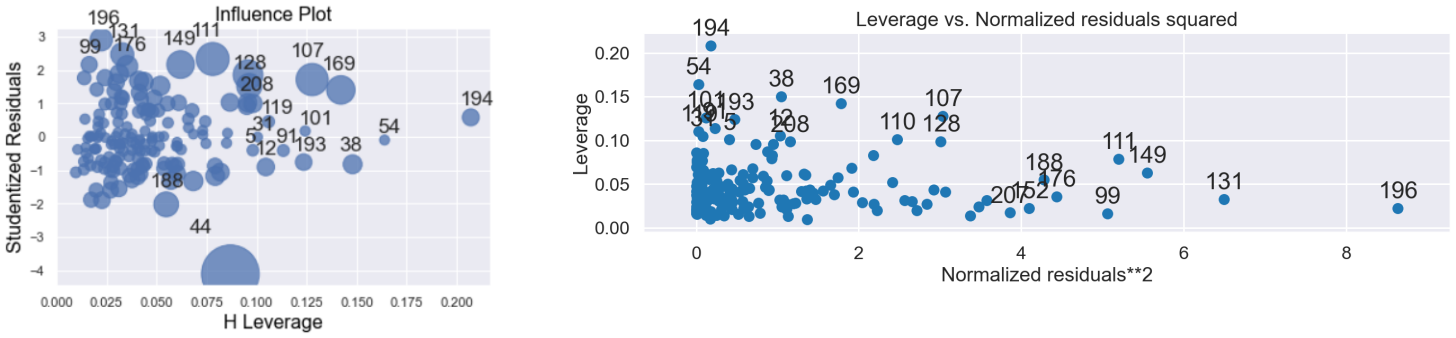


Figure 5. Influence, Leverage Plots

There are 3 other data points that are suspected to be outliers (idx.196 - Trinidad and Tobago, idx.131 - Mongolia, and idx.176 - South Africa). As opposed to Congo, they have higher CO_2 emission per capita than usual given their percentage of GDP produced by the industrial and manufacturing sector. Similar to Congo, it is likely caused by more complex data like geographic condition not included in our dataset. However since those are not as strongly "outlier" as Congo, they are not removed from the dataset.

The Residual Normal Probability Plot looks roughly normal, and all Residuals *vs.* Predictor plots seem to roughly satisfy the assumptions of independence of errors and constant variance. Therefore, multiple linear regression is a valid model to use.

R-squared is 0.881 and Adjusted R-squared is 0.874, meaning there are likely some insignificant predictors in our initial model (see *Dropped Variables Part 2)*. However, it also indicates that the accuracy of the model is quite high.

Dropped Variables (Part 2)

Starting with the full set of predictors, insignificant predictors are dropped, starting from the ones with highest p>|t| value. **Percentage of forested land (pland_forest)** was first dropped with p>|t|=0.66. **Percentage of GDP Growth (pgdp_growth,** Yeo-Johnson Transformed) followed with p>|t|=0.317. The remaining predictors stay within a significance level of 0.25 (we chose a relaxed cutoff for retaining variables in our model, as environmental/economic phenomena are highly complex). Adjusted R-squared remained the same when these predictors were removed.

Removing any of the predictors with the next two highest p>|t| values, started to reduce the Adjusted R-squared value. Therefore, we kept these lower-significance variables in our finalized model. In interpreting our model parameters below, we employed a significance cutoff of 0.05.

Finalized Model

The model is then re-examined against the linear regression assumptions:

1. Individual predictors are linear to the dependent variable (as shown in simple OLS linear regression examinations)
2. The residual distribution is roughly normal. (Fig.5)

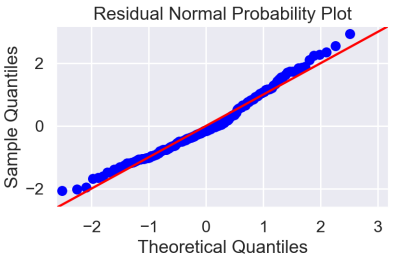


Figure 5. Residual Normal Probability Plot

3. The residuals seem random with no clear pattern, satisfying the independence of errors assumption. (Fig.6)
4. The residuals roughly formed a horizontal band around 0. There are always about 3 data points that slightly stretch the upper limit of the residuals vs predictor plot. After investigation, we identified that those are the other 3 suspected outliers that we did not remove from our dataset. Those are included in the dataset to make the regression model more inclusive. Excluding those 3 outliers, the other points perfectly satisfy the constant variance assumption. We suggest that our model still satisfy this assumption. (Fig.6)

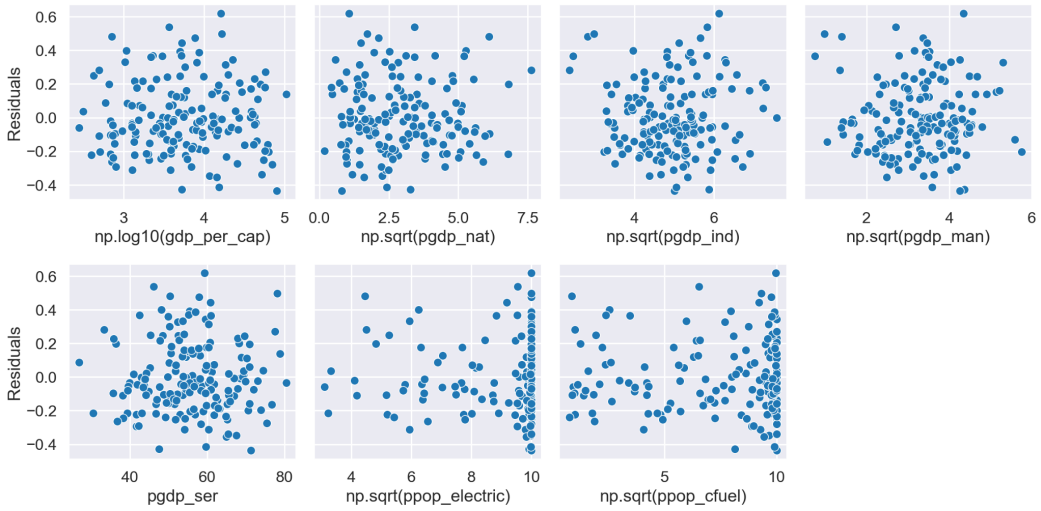


Figure 6. Residuals vs Predictor Plots

The modified model still satisfies the linear regression assumptions. This is therefore our final model (Table 3, Supplementary Table 1). The R-square, 0.88, is very close to the **Adjusted**

R-squared, 0.874, meaning there is likely no overfitting. This also means that the model can explain approx. 87.4% of the response.

Table 3. Summary of key results from multiple OLS linear regression model of $\log_{10}CO_2$ emissions in kt per capita.

* We used back transformations on transformed y_{tr} and X_{tr} to understand the model results in terms of untransformed y (CO_2 emissions in kt per capita) and X predictors. See Table 1 for list of variable transformations applied in our model, and *Data Preparation: Back-transformation methods*.

X variable	Intercept			Slope		
	OLS Intercept	Default value of y when X_{tr} = 0 (kt CO_2 per capita)*	Value of X when X_{tr} = 0*	**OLS Slope**	Change in y per unit increase in X_{tr} (change in kt CO_2 per capita)*	Change in X when X_{tr} increases by 1 unit*
GDP per capita (current US$)	-5.0714	8.484646e-6	1	0.45	Average change of approx. 183%	1000%
GDP growth (annual %)			Excluded from final formula			
Industry (including construction), value added (% of GDP)			0	0.087	Average change of. approx. 22.1%	$2X_{tr}+1$
Manufacturing, value added (% of GDP)			0	-0.054	Average change of. approx. -11.67%	$2X_{tr}+1$
Agriculture, forestry, and fishing, value added (% of GDP)			0	-0.077	Average change of. approx. -16.25%	$2X_{tr}+1$
Services, value added (% of GDP)			0	-0.0062	Average change of. approx. -1.42%	1 (Not transformed)
Forest area (% of land area)			Excluded from final formula			
Access to clean fuels and technologies for cooking (% of population)			0	0.0396	Average change of. approx. 9.55%	$2X_{tr}+1$
Access to electricity (% of population)			0	0.0812	Average change of. approx. 20.57%	$2X_{tr}+1$

Model Interpretation

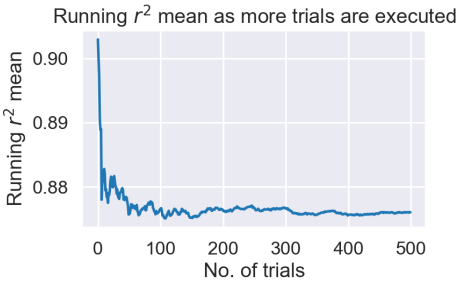
The finalized multiple OLS model showed that by far the strongest contributor to CO_2 emissions per capita was GDP per capita. The model coefficient (slope) for this predictor was 0.45, meaning that a 10-fold increase in GDP per capita is associated with an average increase in CO_2 emissions per capita of 183%. The absolute coefficients of all other variables were at least 4-fold smaller than the coefficient of GDP per capita, demonstrating that they contributed far less to the predicted value of CO_2 per capita.

Accuracy Examination

The final module was created by fitting the full set of data. A different model will be generated if we fit a subset of the dataset. To test the accuracy of the final formula, we:

1. Randomly select a train and test dataset with a 0.7:0.3 ratio
2. Fit the formula
3. Predict CO_2 emission per capita by feeding in test dataset
4. Compute r-squared value using the predicted y and actual y of the test dataset.
5. Repeat this 500 times.

The following plot demonstrates the fluctuation of the running R-squared mean as we execute the 500 trials:



Repeating this procedure multiple times suggests that the R-squared mean often started approx. between 0.84 to 0.9. The running R-squared mean then quickly moves closer to approx. 0.877 and becomes relatively stable near it after around 150th to 300th trials.

The observation seems to suggest that the formula actually has a slightly higher R-squared (0.877) than the adjusted R-squared from our final model (0.874). Since the training set is 70% of the full dataset, the chance that all 3 retained outliers are located within the training set is higher than their chance to be in the testing set. This may mean that the 3 outliers may not be too influential to the model, and hence the accuracy might have slightly increased when the 3 outliers are excluded from the testing set.

Conclusion

From the CO_2 emissions per capita by country (Figure 2A), we expect that developed countries or countries that have a high GDP (i.e. rich countries) tend to have a higher level of CO_2 emissions per capita. Our model is able to prove our expectation where the GDP per capita in each country has the strongest relationship to the CO_2 emissions per capita and the relationship is strongly positive. Industry (% GDP), Clean fuel technology (% population) and Electricity (% population) also have a positive association with the CO_2 emissions per capita. It indicates that if a country has a high industry income or has a huge amount of clean fuel technology and electricity resources, it also has a high level of CO_2 emissions per capita. On the other hand, Manufacturing, Agriculture, and Services (% GDP) has a negative association with the CO_2 emissions per capita, indicating that when a country has a high manufacturing, agriculture, and services income, it has a low level of CO_2 emissions per capita. The positive association of CO_2 per capita in clean fuels goes against our initial expectations, since we expect that if a country has advanced clean fuel technology, it would have a low level of CO_2 emissions per capita.

Looking at the multiple linear regression summary output (in Appendix), we can conclude that the association between the response and each term in the model except Services (% GDP) is statistically significant, given the p-value for each term is less than our significance level (0.05). The adjusted R-sq tells us that 87.4% of variation is explained by only the independent variables that actually affect the dependent variable. The p-value for the F-statistic is less than the significance level of 0.05, which means we can reject the null hypothesis that an intercept-only model is better, *i.e.* all independent variables are significant as a whole. The multiple regression regression is a better model than our best simple OLS regression model (CO_2 per capita *vs.* GDP per capita) since the adjusted R-square increased by 7.5%. This means the multiple regress is able to capture a significant amount of additional information or variation that GDP per capita can't capture by itself.

Furthermore, our model is adequate and meets all assumptions of the regression by checking the residuals *vs.* predicted values, residuals *vs.* X's, and Q-Q plots.

If we want to improve the model, one of the things we can do is to use different methods to deal with the missing values in the predictor variables. Besides removing observations with any missing independent values or fill the missing with the median, we can also try the following methods and see which would give us a better model performance:

- Replace with some constant value outside fixed value range: -999 or -1, etc.
- Replace with median by cluster (imputation using k-NN)
- Imputation Using Multivariate Imputation by Chained Equation (MICE)
    - filling the missing data multiple times
    - Multiple Imputations (MIs) are much better than a single imputation as it measures

the uncertainty of the missing values in a better way and it can handle different variables of different data types (ie., continuous or binary)

- Stochastic regression imputation
    - predict the missing values by regressing it from other related variables in the same dataset plus some random residual value
- Multivariate feature imputation
    - a strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion

Appendix

Supplementary Tables

OLS Regression Results

Dep. Variable:	log10_co2_per_cap	R-squared:	0.880
Model:	OLS	Adj. R-squared:	0.874
Method:	Least Squares	F-statistic:	165.1
Date:	Sat, 10 Apr 2021	Prob (F-statistic):	2.80e-69
Time:	11:20:39	Log-Likelihood:	22.770
No. Observations:	166	AIC:	-29.54
Df Residuals:	158	BIC:	-4.643
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.0714	0.536	-9.456	0.000	-6.131	-4.012
np.log10(gdp_per_cap)	0.4518	0.067	6.723	0.000	0.319	0.585
np.sqrt(pgdp_nat)	-0.0770	0.036	-2.122	0.035	-0.149	-0.005
np.sqrt(pgdp_ind)	0.0867	0.034	2.581	0.011	0.020	0.153
np.sqrt(pgdp_man)	-0.0539	0.021	-2.543	0.012	-0.096	-0.012
pgdp_ser	-0.0062	0.004	-1.586	0.115	-0.014	0.002
np.sqrt(ppop_electric)	0.0812	0.020	4.124	0.000	0.042	0.120
np.sqrt(ppop_cfuel)	0.0396	0.014	2.886	0.004	0.012	0.067

Omnibus:	7.068	Durbin-Watson:	2.091
Prob(Omnibus):	0.029	Jarque-Bera (JB):	7.328
Skew:	0.510	Prob(JB):	0.0256
Kurtosis:	2.869	Cond. No.	1.89e+03

Supplementary Table 1. OLS Regression Results of finalized multivariate model.